NVIDIA

Datacenter용 Mellanox 네트워킹 플랫폼 소개

베이넥스

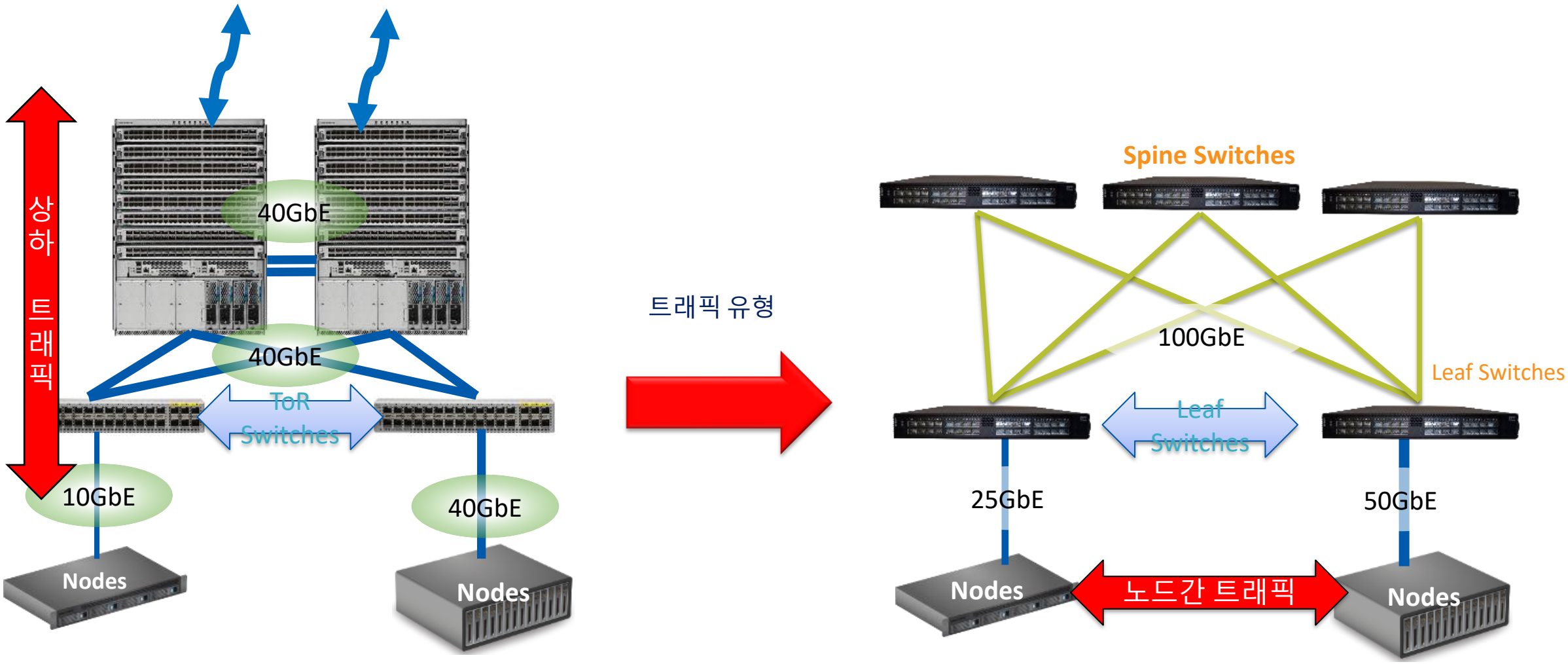
v01 2021.10

BayNex

Datacenter

- Mellanox 제품군의 필요성

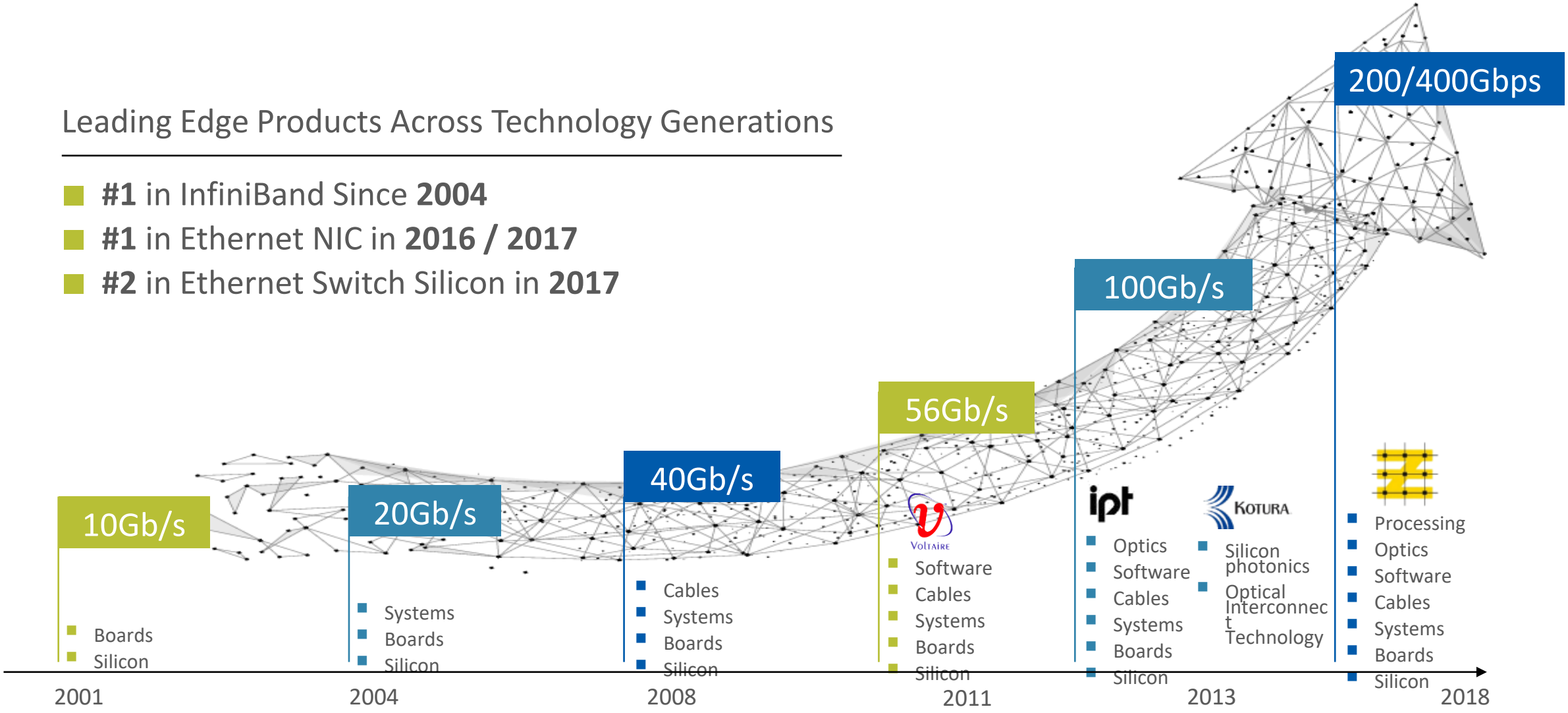
Datacenter 트래픽 유형변화 : 고속의 East-West 트래픽



고속 트래픽 영역의 선도적/지속적 기술주도

Leading Edge Products Across Technology Generations

- **#1** in InfiniBand Since **2004**
- **#1** in Ethernet NIC in **2016 / 2017**
- **#2** in Ethernet Switch Silicon in **2017**



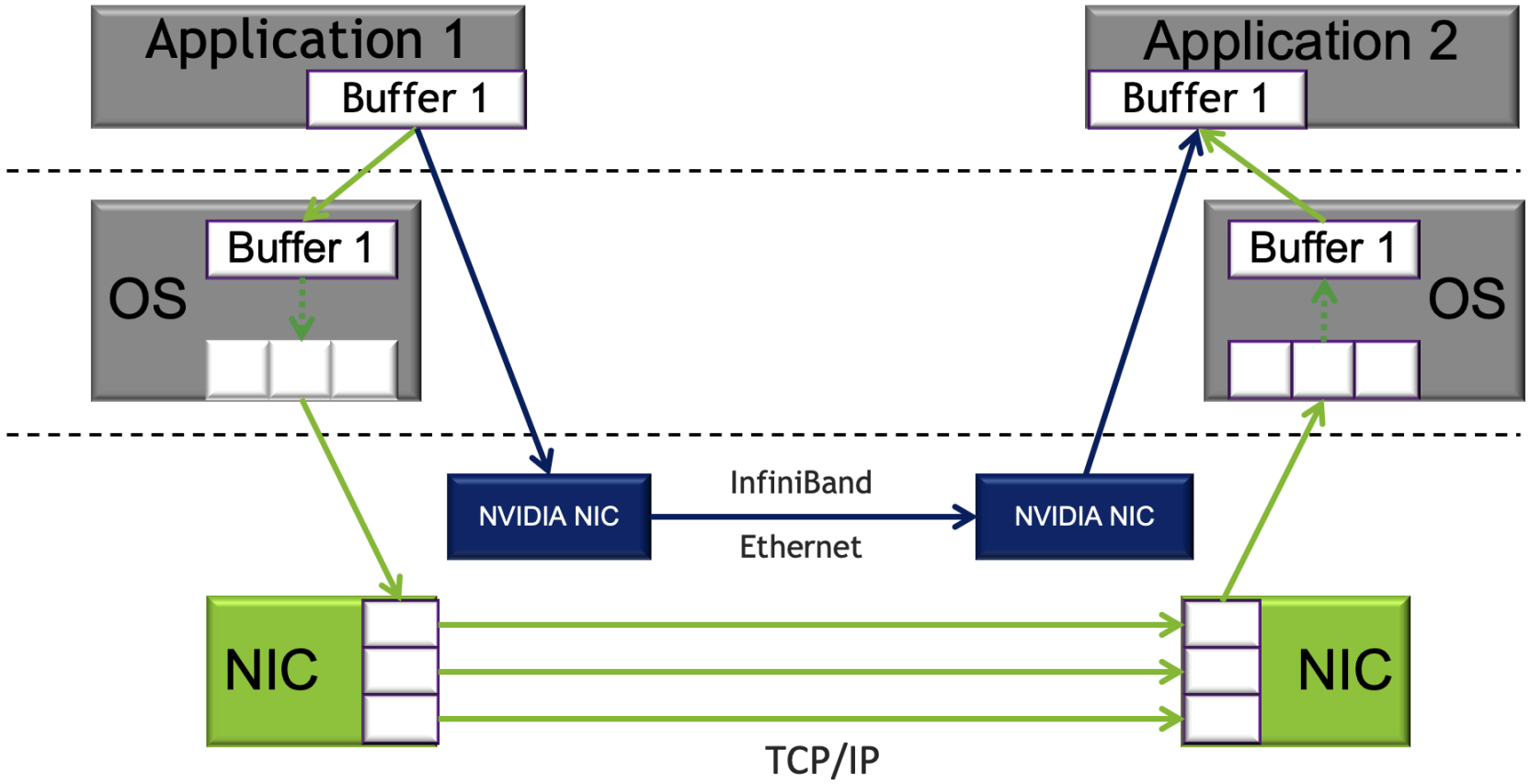
Mellanox

- Infiniband 제품 특징

Infiniband : Key Features

Infiniband : RDMA

Remote Direct Memory Access

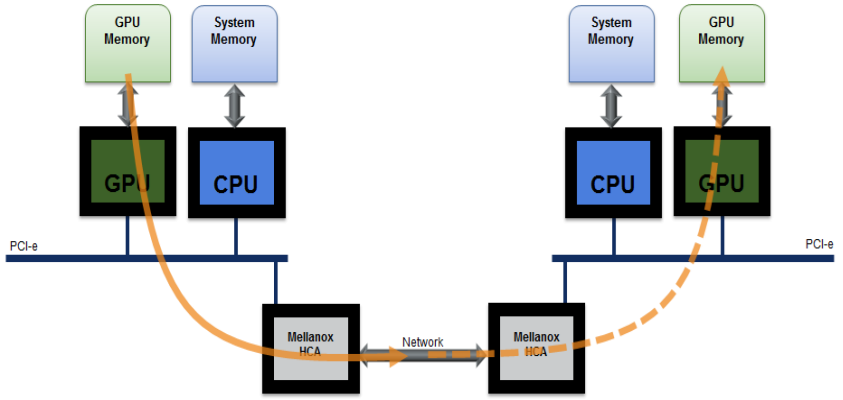


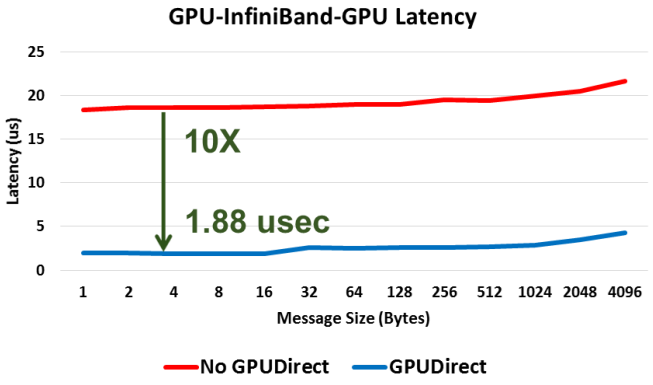
Infiniband : GPUDirect™ RDMA

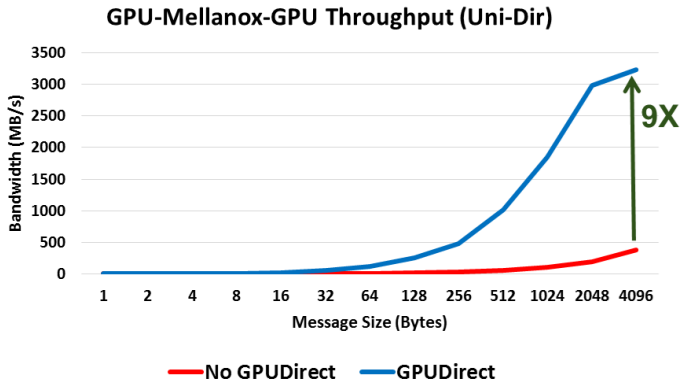
10X higher Performance

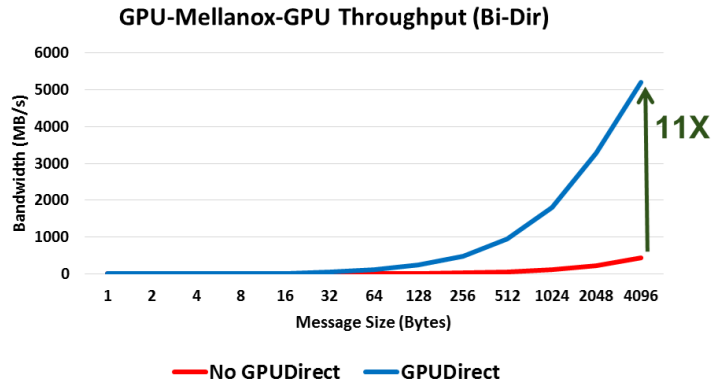
- Accelerates HPC and Deep Learning performance
- Lowest communication latency for GPUs

GPUDirect™ RDMA







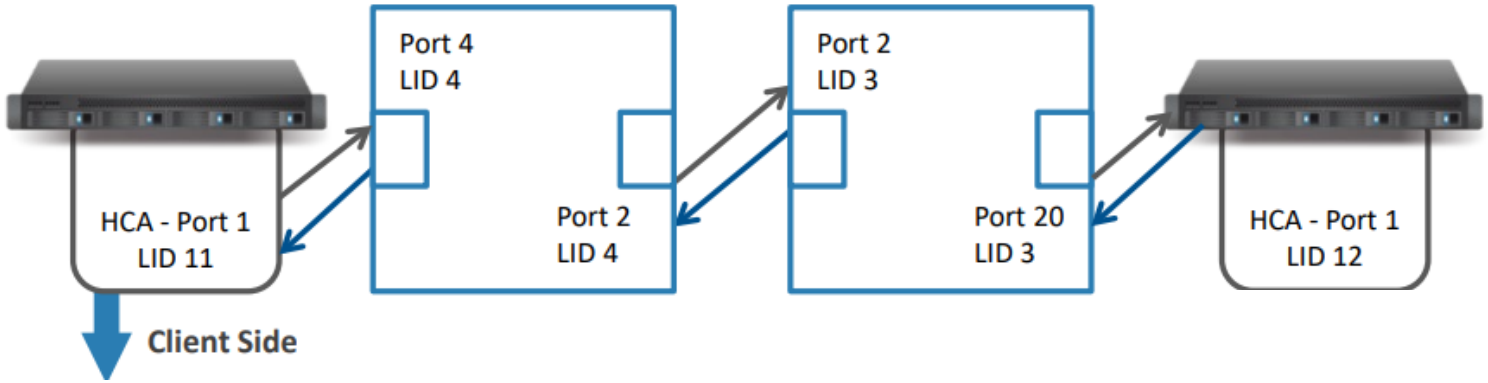


Infiniband 통신1 - IP주소 (IPoIB방식)

```
[root@mtlacad02 ~]# ifconfig ib0
Ifconfig uses the ioctl access method to get the full address information, which limits hardware
addresses to 8 bytes.
Because Infiniband address has 20 bytes, only the first 8 bytes are displayed correctly.
Ifconfig is obsolete! For replacement check ip.
ib0       Link encap:InfiniBand  HWaddr A0:00:02:08:FE:80:00:00:00:00:00:00:00:00:00:00:00:00:00:00
          inet addr:192.168.51.2  Bcast:192.168.51.255  Mask:255.255.255.0
          inet6 addr: fe80::f652:1403:33:d1d1/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:2044  Metric:1
          RX packets:0 errors:0 dropped:0 overruns:0 frame:0
          TX packets:6 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1024
          RX bytes:0 (0.0 b)  TX bytes:456 (456.0 b)
```

Infiniband 통신2 - LID주소 (Infiniband고유방식)

```
[mtlacad05@mtlacad05 ~]$ ibstatus
Infiniband device 'mlx5_0' port 1 status:
        default gid:     fe80:0000:0000:0000:7cfe:9003:005d:7b2e
        base lid:        0x5
        sm lid:          0x12
        state:           4: ACTIVE
        phys state:      5: LinkUp
        rate:            56 Gb/sec (4X FDR)
        link_layer:      InfiniBand
```



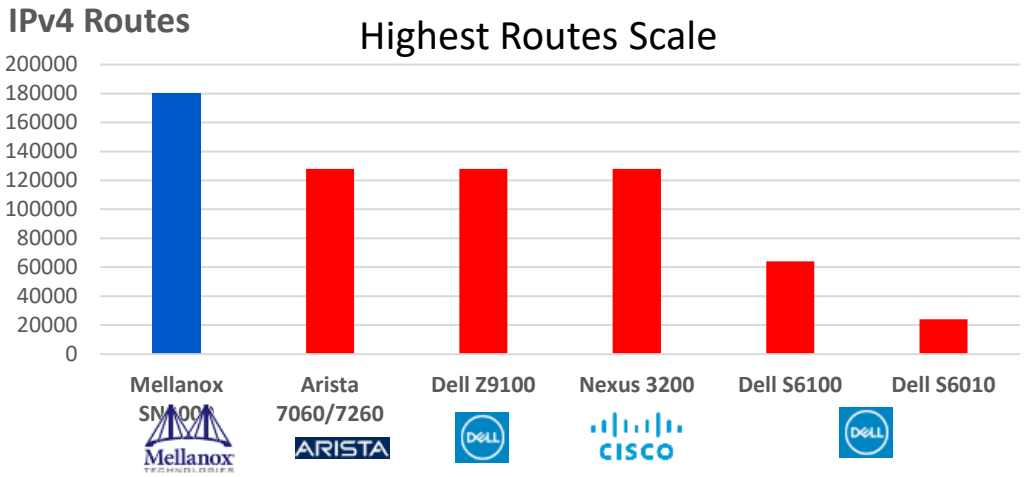

```
[root@ib-cert-sv02 ~]# ibping -L 12
Pong from ib-cert-sv01.lab.mtl.com (Lid 12): time 0.141 ms
Pong from ib-cert-sv01.lab.mtl.com (Lid 12): time 0.085 ms
Pong from ib-cert-sv01.lab.mtl.com (Lid 12): time 0.082 ms
Pong from ib-cert-sv01.lab.mtl.com (Lid 12): time 0.056 ms
Pong from ib-cert-sv01.lab.mtl.com (Lid 12): time 0.070 ms
```

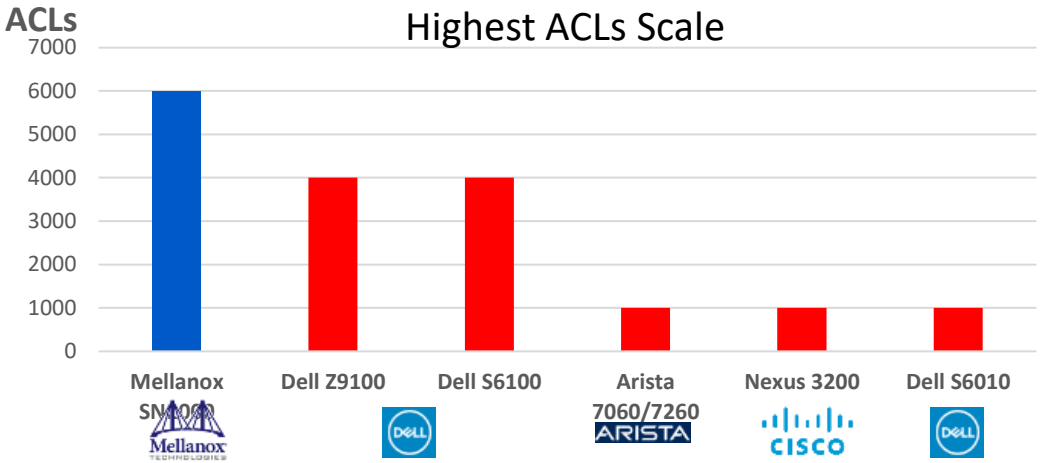
Mellanox

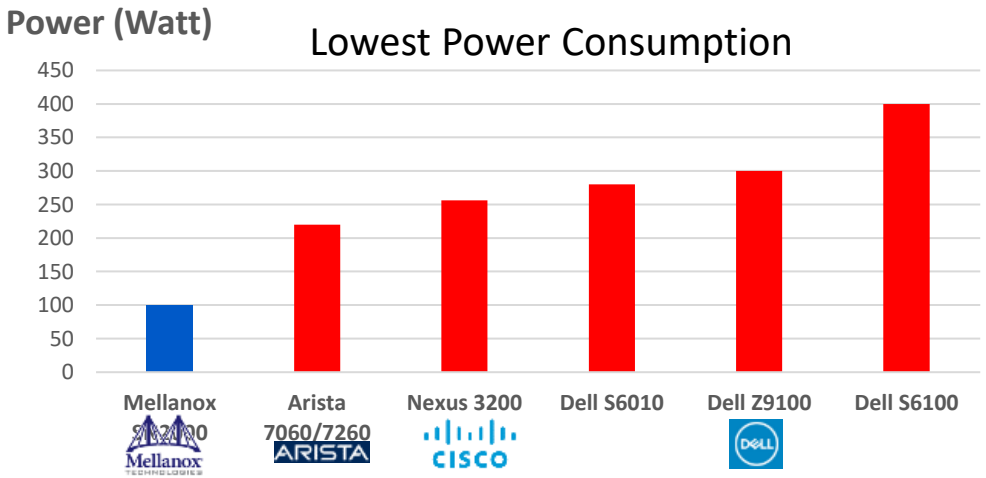
- Ethernet 제품 특징

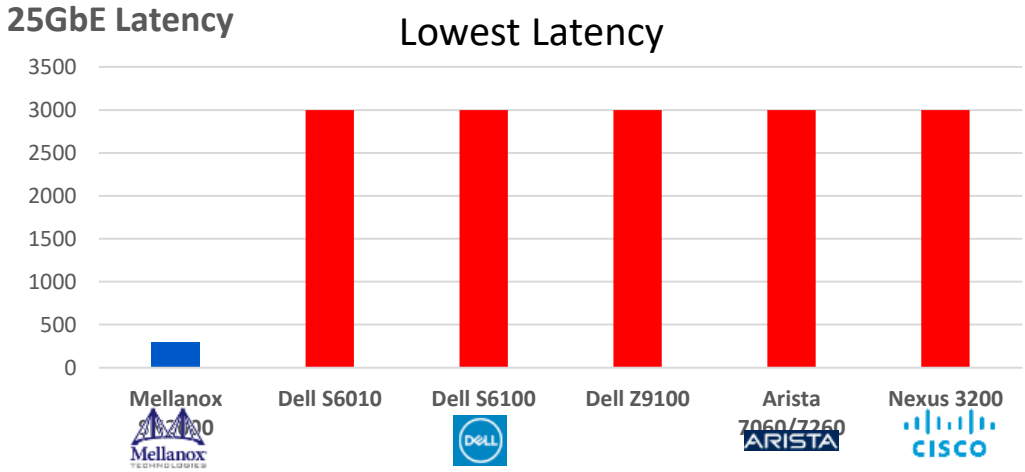
이더넷 : Spectrum vs BROADCOM

2016.10









이더넷 : 고속 cut-through 방식

Cut-Through vs. Store and Forward

- Mixed setting on port speeds
  - 10/25/40/50Gig Downlinks
  - 100Gig Uplink

Downlink to Downlink

Full Cut-Through

Uplink to Downlink

Full Cut-Through

Downlink to Uplink

Smart Store and Forward

이더넷 : ROCE

RDMA

(Remote Direct

Memory Access)

Without RDMA

With RDMA

ROCE

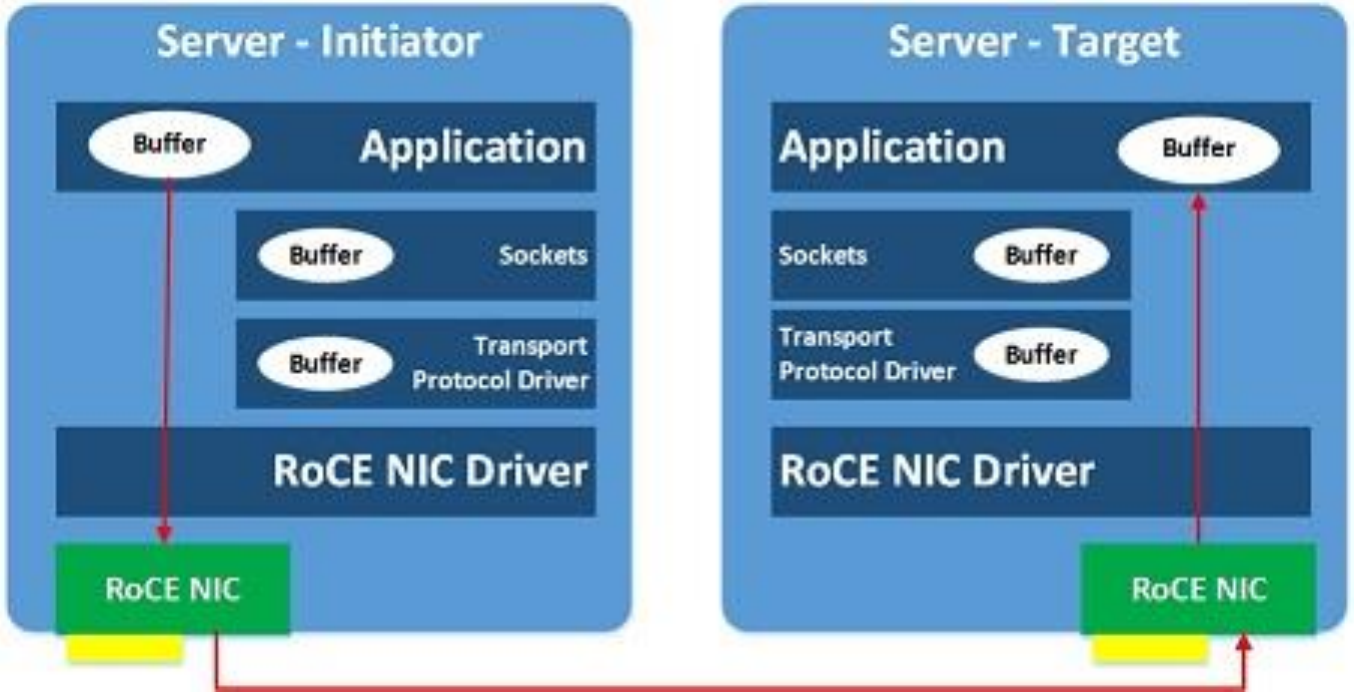
(RDMA over

Converged Ethernet)

Microsoft Azure

Alibaba.com

Dennis Cai
Chief Architect



Mellanox

- 제품군

Mellanox 제품군 / 용도

Infiniband
스위치

or

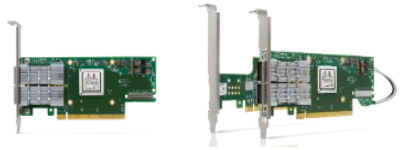
이더넷
스위치

Mellanox 제품군

Adapters

ConnectX-6

IB 카드 : HDR(200Gb/s)
이더넷 카드 : 200GbE
(ConnectX6칩)



IB Switch

SwitchIB-2 Quantum

IB EDR 스위치 (SwitchIB칩)
IB HDR 스위치 (Quantum칩)

이더넷 Switch

Spectrum™

이더넷 100G/200G 스위치 (Spectrum칩)

Interconnect

LinkX™

케이블, 트랜시버 (IB용/이더넷용 구분)

SoC

BlueField-2

SoC칩 (ConnectX6칩 + ARM cpu)
및 칩이 장착된 카드

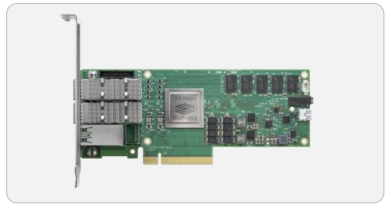
IB (Infiniband) 스위치

제품	SB7800	QM8700	QM8790
			
특징	EDR (100G)	HDR (200G)	HDR (200G) (DGX-A100 SuperPod)
Subnet manage	2000 노드	2000 노드	n/a
포트	EDR 36포트	HDR 40포트	HDR 40포트
Throughput	7.2 Tb/s	16 Tb/s	16 Tb/s
Latency	90 ns	130 ns	130 ns
크기	1U	1U	1U

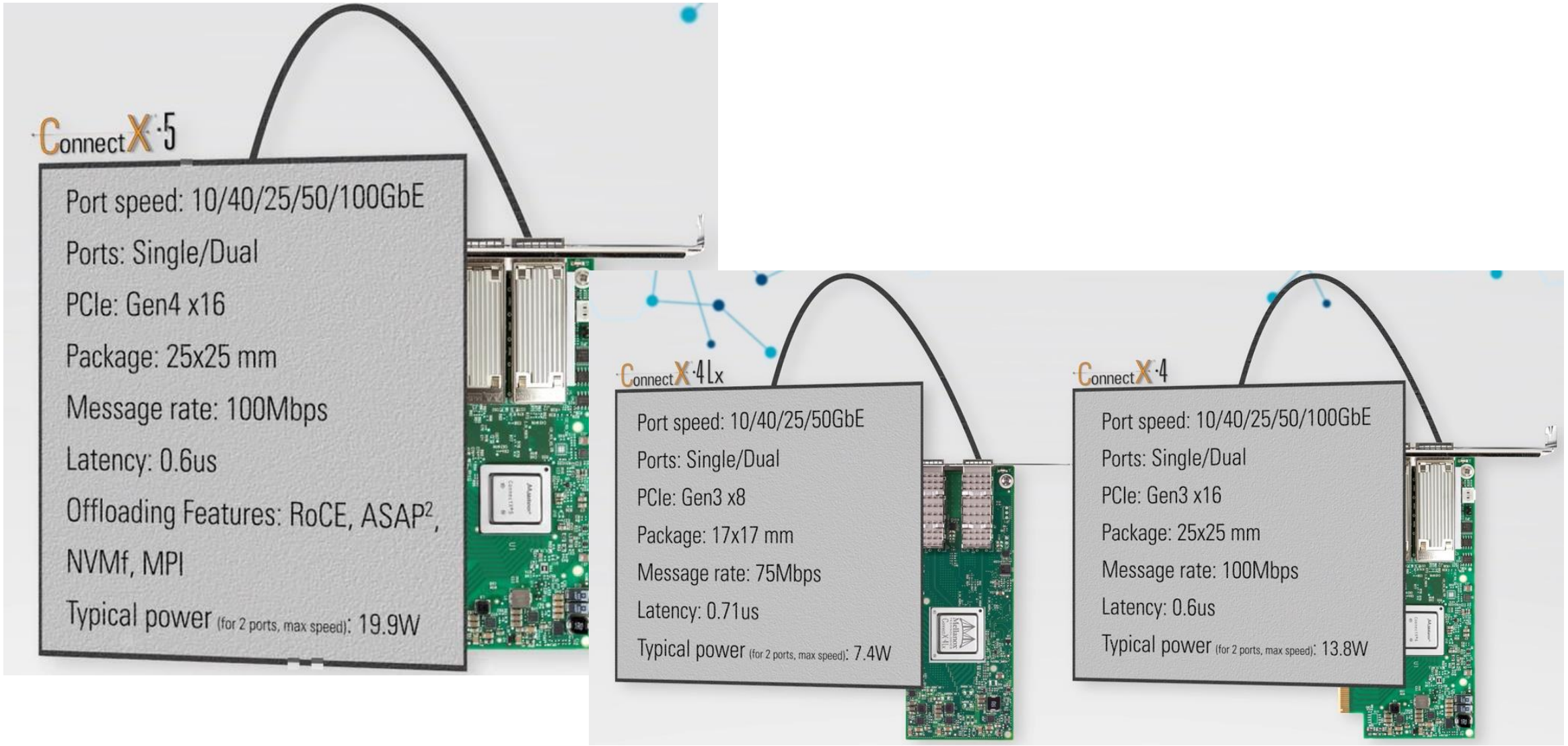
이더넷 스위치

제품	SN2010	SN2100	SN3700c	SN4600c
				
특징	10/25G + 100G (소규모)	100G (소규모)	100G	100G (DGX-A100 SuperPod)
포트	10/25GbE SFP28 18포트 + 100GbE QSFP28 4포트	100GbE QSFP28 16포트	100GbE QSFP28 32포트	100GbE QSFP28 64포트
Throughput	1.7 Tb/s 1.26 B(billion)pps	3.2 Tbps 2.38 B(billion)pps	6.4 Tbps 4.76 B(billion)pps	12.8 Tbps 8.4 B(billion)pps
Latency	300 ns	300 ns	425 ns	425 ns
크기	Half 1U	Half 1U	1U	2U

어댑터

제품	ConnectX-6	BlueField-2	BlueField-3
			NVIDIA
특징	HDR (200Gb/s)	ConnectX-6 + Arm A72 core	ConnectX-7 + Arm A78 core
IB 전용 카드	O	O	-
이더넷 전용 카드	O	O	-
VPI 카드 (IB/이더넷 겸용)	O	O	-
PCI 인터페이스	PCIe Gen3/4 x16	PCIe Gen3/4 x16	PCIe Gen5 x16/32
			(칩 개발 완료)

이전 어댑터 모델


ConnectX-5
Port speed: 10/40/25/50/100GbE
Ports: Single/Dual
PCIe: Gen4 x16
Package: 25x25 mm
Message rate: 100Mbps
Latency: 0.6us
Offloading Features: RoCE, ASAP², NVMf, MPI
Typical power (for 2 ports, max speed): 19.9W
ConnectX-4 Lx
Port speed: 10/40/25/50GbE
Ports: Single/Dual
PCIe: Gen3 x8
Package: 17x17 mm
Message rate: 75Mbps
Latency: 0.71us
Typical power (for 2 ports, max speed): 7.4W
ConnectX-4
Port speed: 10/40/25/50/100GbE
Ports: Single/Dual
PCIe: Gen3 x16
Package: 25x25 mm
Message rate: 100Mbps
Latency: 0.6us
Typical power (for 2 ports, max speed): 13.8W

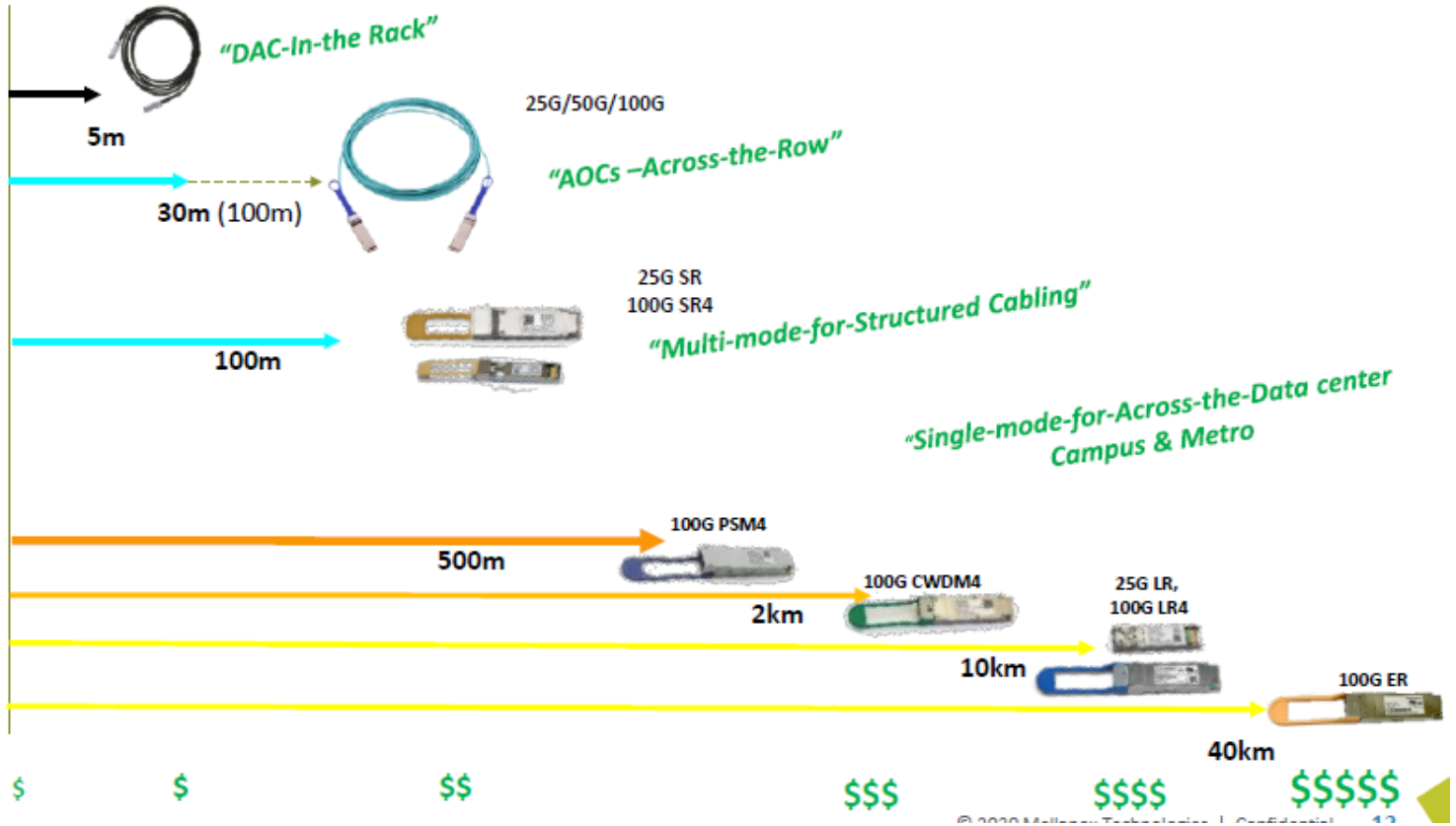
LinkX (케이블,트랜시버)

제품	DAC (Direct Attach Copper)	AOC (Active Optical Cable)	트랜시버
			
특징	단거리용 (<5M)	중거리용 (<100M)	높은 호환성 (타 벤더 장비 연결)
IB 전용	O	O	O
이더넷 전용	O	O	O
IB/이더넷 겸용	일부 FDR(40G)	없음	없음

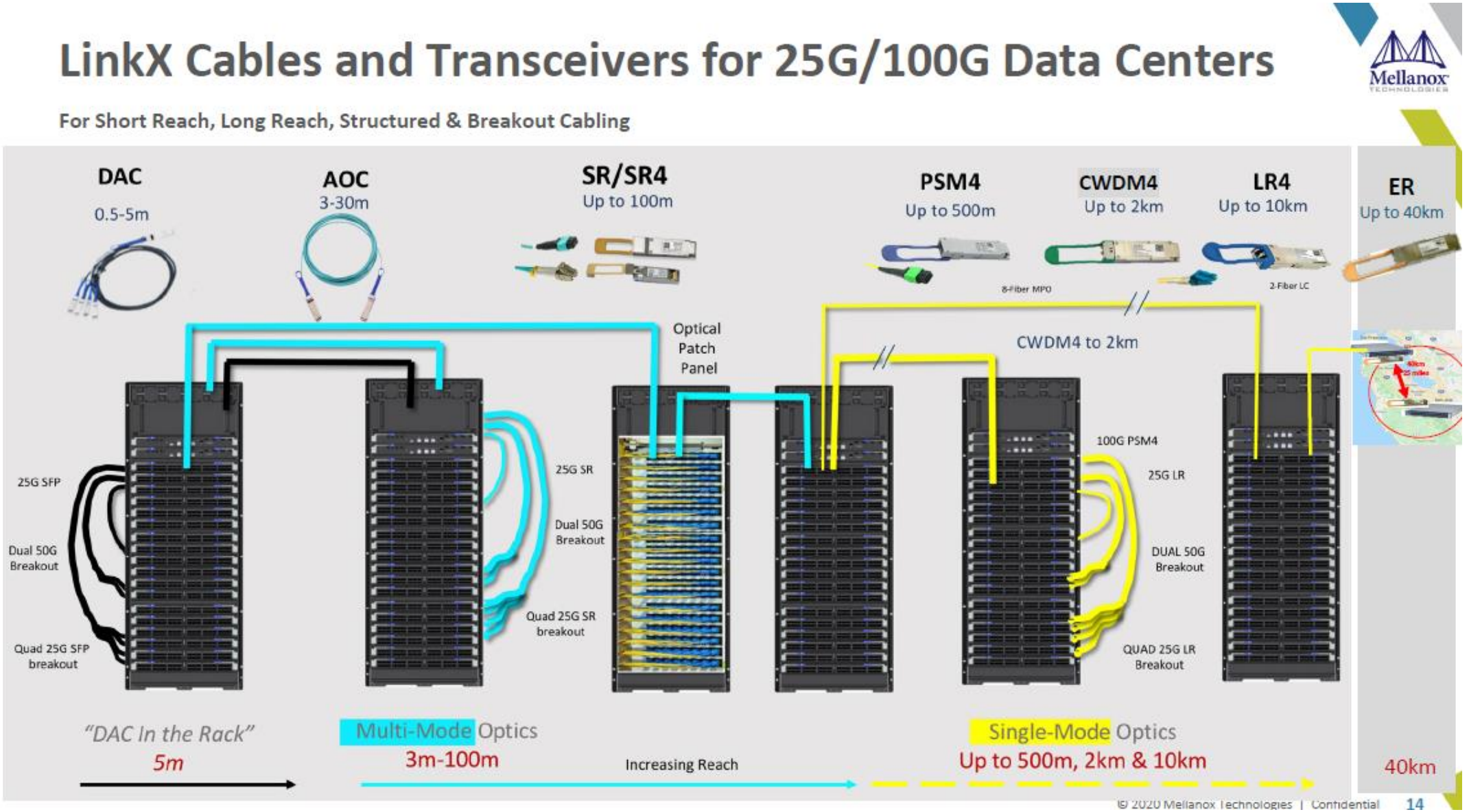
거리별 사용 제품

Interconnects are Cost Optimized for Each Reach

- DAC Copper Cables
- AOCs
- Multi-Mode Transceivers
  - Up to 100m
- Single-Mode Transceivers
  - 500m
  - 2km
  - 10km
  - 40km



거리별 사용 제품


LinkX Cables and Transceivers for 25G/100G Data Centers
For Short Reach, Long Reach, Structured & Breakout Cabling
Mellanox
DAC
0.5-5m
AOC
3-30m
SR/SR4
Up to 100m
PSM4
Up to 500m
CWDM4
Up to 2km
LR4
Up to 10km
ER
Up to 40km
8-Fiber MPO
2-Fiber LC
Optical Patch Panel
CWDM4 to 2km
25G SFP
Dual 50G Breakout
Quad 25G SFP breakout
25G SR
Dual 50G Breakout
Quad 25G SR breakout
100G PSM4
25G LR
DUAL 50G Breakout
QUAD 25G LR Breakout
"DAC In the Rack"
5m
Multi-Mode Optics
3m-100m
Increasing Reach
Single-Mode Optics
Up to 500m, 2km & 10km
40km

트랜시버

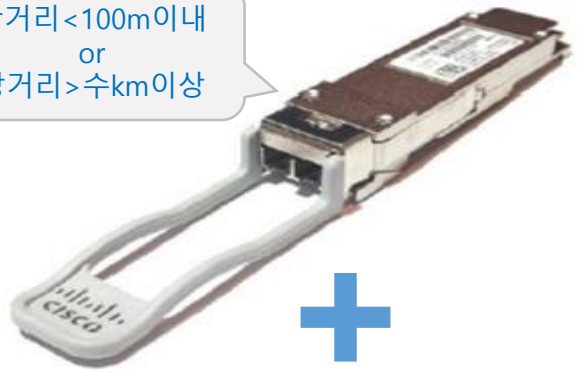
10G/25G
SFP, LC타입 트랜시버

+

LC-LC
케이블

40G/100G
QSFP, LC타입 트랜시버

단거리<100m이내
or
장거리>수km이상

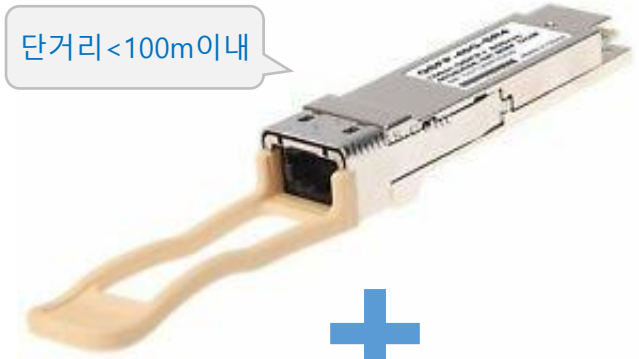


+

LC-LC
케이블

40G/100G/200G, FDR/EDR/HDR
QSFP, MPO타입 트랜시버

단거리<100m이내



+

MPO-MPO
케이블

NVIDIA

감사합니다

BayNex